

Maximal Unbordered Factors of Random Strings*

Patrick Hagge Cording^{† 1} and Mathias Bæk Tejs Knudsen^{‡ 2}

¹DTU Compute, Technical University of Denmark, phaco@dtu.dk

²Department of Computer Science, University of Copenhagen,
mathias@tejs.dk

April 17, 2017

Abstract

A border of a string is a non-empty prefix of the string that is also a suffix of the string, and a string is unbordered if it has no border. Loptev, Kucherov, and Starikovskaya [CPM '15] conjectured the following: If we pick a string of length n from a fixed alphabet uniformly at random, then the expected length of the maximal unbordered factor is $n - O(1)$. We prove that this conjecture is true by proving that the expected value is in fact $n - \Theta(\sigma^{-1})$, where σ is the size of the alphabet. We discuss some of the consequences of this theorem.

1 Introduction

A string S is a finite sequence of n characters from an alphabet Σ of size σ . $S[i, j]$, $1 \leq i \leq j \leq n$, is the sequence of characters of S starting in i and j , both indices included. We denote $S[i, j]$ a *factor* of S . The factor $S[1, j]$ is a prefix of S and $S[i, n]$ is a suffix. A *border* of a string is a non-empty prefix of the string that is also a suffix of the string. If $S = \alpha\beta = \lambda\alpha$, for non-empty strings β and λ , then α is a border of S with length $|\alpha|$. The

*The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-46049-9_9

[†]Supported by the Danish Research Council under the Sapere Aude Program (DFR 4005-00267).

[‡]Research partly supported by Advanced Grant DFF-0602-02499B from the Danish Council for Independent Research under the Sapere Aude research career programme and by the FNU project AlgoDisc - Discrete Mathematics, Algorithms, and Data Structures

maximal border of S is the longest border among all borders of S . S is unbordered if it does not have a border. The maximal unbordered factor is the longest factor that does not have a border. A string is periodic if it can be written as $S = \alpha^k \alpha'$, where α^k is the string α repeated $k > 0$ times and α' is a prefix of α .

Borders were first studied by Ehrenfeucht and Silberger [2] with emphasis on the relationship between the maximal unbordered factor of a string and its minimal period. This relationship has since received more attention in the literature [1, 4, 5].

Loptev, Kucharov, and Starikovskaya [10] prove that for $\sigma \geq 2$ the expected length of the maximal unbordered factor is at least $n(1 - \xi(\sigma) \cdot \sigma^{-4}) + O(1)$, where $\xi(\sigma)$ converges to 2 as σ grows. When $\sigma \geq 5$ and n is sufficiently large this implies that the expected length of the maximal unbordered factor is at least $0.99n$. Supported by experimental results, the authors of [10] conjectured that the expected length of maximal unbordered factor is $n - O(1)$. We prove that this conjecture is true and obtain the following theorem.

Theorem 1. *Let S be a string of length n , where each character is chosen i.i.d. uniformly from an alphabet A of size $\sigma \geq 2$. The expected length of the maximal unbordered factor is $n - O(\sigma^{-1})$.*

The problem of computing the maximal unbordered factor of a string has been studied by Loptev et al. [10] and Gawrychowski et al. [3], who give algorithms with average-case running times $O(\frac{n^2}{\sigma^4} + n)$ and $O(n \log n)$, respectively. It can be decided in $O(n)$ time if a string of length n has a border by computing the *border array* (also known as the failure function, made famous by the KMP pattern matching algorithm [7, 11]). Entry i of the border array B of a string S contains the length of the maximal border of the prefix $S[1, i]$. If $B[n] = 0$ then S is unbordered. Let B_j be the border array for the suffix $S[j, n]$. If $B_j[i] = 0$ it means that the factor $S[j, i]$ is unbordered. Computing B_j for $j = 1 \dots n$ and scanning these to find the maximal unbordered factor of S takes $O(n^2)$ time. As mentioned in [10], we can compute the B_j 's in decreasing order of the suffix length and stop the algorithm once $n - j$ is smaller than the currently longest unbordered factor and obtain an algorithm with average-case running time $O((n - \mu + 1)n)$ where μ is the expected length of the maximal unbordered factor. With our new bound on the expected length of the maximal unbordered factor, we therefore get the following corollary.

Corollary 1. *There is an algorithm with average-case running time $O(n)$ that finds the maximal unbordered factor.*

This improves the previously best known average-case bounds for finding the maximal unbordered factor of a string.

Related work. The worst-case running time of the above mentioned algorithm is still $O(n^2)$. Gawrychowski et al. [3] give an algorithm with worst-case running time $O(n^{1.5})$.

Holub and Shallit [6] investigated the expected length of the maximal border of a random word.

Data structures for answering a border query have also been developed. A border query takes two indices i and j and the answer is the maximal border of the factor $S[i, j]$. Kociumaka et al. [8] show several time-space trade-offs for this problem. For one of these, their data structure can answer border queries in $O(\log^{1+\epsilon} n)$ time and uses $O(n)$ space. Kociumaka et al. [9] improved this to $O(1)$ time for answering border queries while using $O(n)$ space.

2 The Proof of Theorem 1

Fix A and $\sigma \geq 2$. Let X_n be the expected length of the maximal unbordered factor of a random string of length n . We define $X_0 = 0$, and we let $Y_n = n - X_n$. We prove in the following that $Y_n \leq c$, where c is given by:

$$c = \frac{2\sigma}{(\sigma - 1)^2(1 - \sigma^{-1} - \sigma^{-2})}$$

Since $c \leq \frac{32}{\sigma}$ this will prove the theorem. This follows from $\sigma \geq 2$ and the following calculation:

$$c = \frac{2}{\sigma(1 - \sigma^{-1})^2(1 - \sigma^{-1} - \sigma^{-2})} \leq \frac{2}{\sigma(1 - 2^{-1})^2(1 - 2^{-1} - 2^{-2})} = \frac{32}{\sigma}$$

We will prove the claim by induction on n . By definition this is true whenever $n \leq 1$. So fix some n and assume that $Y_m \leq c$ for all $m < n$.

Let S be a random string of length n . Let $f = f(S)$ be the smallest positive integer $< n$ such that $S[1, f] = S[n - f + 1, n]$. If no such integer exists we let $f = 0$. We note that if $f > 0$ then $f \leq \frac{n}{2}$, since if $f > \frac{n}{2}$ then $f' = 2f - n$ satisfies $S[1, f'] = S[n - f' + 1, n]$ as well and $f' < f$ which is impossible. Let $L = L(S)$ be the length of the maximal unbordered factor of S . Then:

$$Y_n = n - X_n = n - E(L) = \sum_{\ell=0}^{\lfloor n/2 \rfloor} P(f = \ell)(n - E(L \mid f = \ell)) \quad (1)$$

If $1 \leq \ell < \frac{n}{2}$ then $S[\ell + 1, n - \ell]$ is independent of the event $f = \ell$, since $f = \ell$ is determined by $S[1, \ell]$ and $S[n - \ell + 1, n]$. The longest unbordered factor in $S[\ell + 1, n - \ell]$ is also an unbordered factor in S and hence for $\ell < \frac{n}{2}$:

$$E(L \mid f = \ell) \geq X_{n-2\ell} \quad (2)$$

If n is odd (2) holds for all integers $\ell \in \{1, 2, \dots, \lfloor \frac{n}{2} \rfloor\}$. If n is even we see that if $\ell = \lfloor \frac{n}{2} \rfloor$ the right hand side of (2) is 0 and hence it also holds for all integers $\{1, 2, \dots, \lfloor \frac{n}{2} \rfloor\}$. If $f = 0$, then S is an unbordered factor, and therefore $E(L \mid f = 0) = n$. So we can use this observation together with the inequality (2) to upper bound Y_n in (1) by:

$$Y_n \leq \sum_{\ell=1}^{\lfloor n/2 \rfloor} P(f = \ell)(n - X_{n-2\ell}) = \sum_{\ell=1}^{\lfloor n/2 \rfloor} 2\ell P(f = \ell) + \sum_{\ell=1}^{\lfloor n/2 \rfloor} P(f = \ell)Y_{n-2\ell} \quad (3)$$

Nielsen [12] proved the following lower bound on the probability that S is unbordered. Since S is unbordered iff $f = 0$ we get:

Theorem 2 (Nielsen [12]).

$$P(f = 0) \geq 1 - \sigma^{-1} - \sigma^{-2}$$

Using Theorem 2 together with the fact that $Y_{n-2\ell} \leq c$ we get:

$$\sum_{\ell=1}^{\lfloor n/2 \rfloor} P(f = \ell)Y_{n-2\ell} \leq c \sum_{\ell=1}^{\lfloor n/2 \rfloor} P(f = \ell) = c(1 - P(f = 0)) \leq c(\sigma^{-1} + \sigma^{-2}) \quad (4)$$

If $f = \ell$ then $S[1, \ell] = S[n - \ell + 1, n]$. After fixing $S[1, \ell]$ there are σ^ℓ ways to choose $S[n - \ell + 1, n]$ and hence $P(f = \ell) \leq \sigma^{-\ell}$. Using this we get:

$$\sum_{\ell=1}^{\lfloor n/2 \rfloor} 2\ell P(f = \ell) \leq \sum_{\ell=1}^{\infty} 2\ell \sigma^{-\ell} = \frac{2\sigma}{(\sigma - 1)^2} \quad (5)$$

Inserting (4) and (5) into (3) gives:

$$Y_n \leq \frac{2\sigma}{(\sigma - 1)^2} + c(\sigma^{-1} + \sigma^{-2}) = c$$

which finishes the induction and the proof. \square

References

- [1] J.-P. Duval. Relationship between the period of a finite word and the length of its unbordered segments. *Discrete Mathematics*, 40(1):31–44, 1982.
- [2] A. Ehrenfeucht and D. Silberger. Periodicity and unbordered segments of words. *Discrete Mathematics*, 26(2):101–109, 1979.
- [3] P. Gawrychowski, G. Kucherov, B. Sach, and T. Starikovskaya. Computing the longest unbordered substring. In *String Processing and Information Retrieval*, pages 246–257. Springer, 2015.
- [4] T. Harju and D. Nowotka. Periodicity and unbordered words: A proof of the extended Duval conjecture. *Journal of the ACM (JACM)*, 54(4):20, 2007.
- [5] Š. Holub and D. Nowotka. The Ehrenfeucht–Silberger problem. *Journal of Combinatorial Theory, Series A*, 119(3):668–682, 2012.
- [6] Š. Holub and J. Shallit. Periods and borders of random words. In *33rd Symposium on Theoretical Aspects of Computer Science*, 2016.
- [7] D. E. Knuth, J. H. Morris, Jr, and V. R. Pratt. Fast pattern matching in strings. *SIAM journal on computing*, 6(2):323–350, 1977.
- [8] T. Kociumaka, J. Radoszewski, W. Rytter, and T. Waleń. Efficient data structures for the factor periodicity problem. In *String Processing and Information Retrieval*, pages 284–294. Springer, 2012.
- [9] T. Kociumaka, J. Radoszewski, W. Rytter, and T. Waleń. Internal pattern matching queries in a text and applications. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 532–551. SIAM, 2015.
- [10] A. Loptev, G. Kucherov, and T. Starikovskaya. On maximal unbordered factors. In *Combinatorial Pattern Matching*, pages 343–354. Springer, 2015.
- [11] J. H. Morris, Jr and V. R. Pratt. *A linear pattern-matching algorithm*. 1970.
- [12] P. T. Nielsen. A note on bifix-free sequences (corresp.). *IEEE Trans. Information Theory*, 19(5):704–706, 1973.